

Best Available Copy

Appl. No. 09/844,730
 Amdt. dated July 19, 2005
 Reply to Notice of Allowance of July 29, 2005

Amendments to the Specification:

Please amend the title as follows:

Distributed Data Clustering System And Method-And-System.

Please amend paragraph 94 to correct the symbols as follows:

[0094] $\delta_k(x) = 1$ if x is closest to m_k , otherwise $\delta_k(x) = 0$ (resolve ties arbitrarily). The summation of these functions over a data set (see (3) and (4)) residing on the l^{th} unit gives the count, $n_{k,l}$, first moment, $\sum_{k,l}$, and the second moment, $s_{k,l}$, of the clusters. The vector $\{n_{k,l}, \sum_{k,l}, s_{k,l} \mid k=1, \dots, K\}$, has dimensionality $2 \cdot K + K \cdot \text{dim}$, which is the size of the SS that have to be communicated between the Integrator and each computing unit.

Please amend paragraph 95 to correct the symbols as follows:

[0095] The set of SS presented here is more than sufficient for the simple version of K-Means algorithm. The aggregated quantity, $\sum_k s_{k,l}$, could be sent instead of the individual $s_{k,l}$. But there are other variations of K-Means performance functions that require individual $s_{k,l}$ for evaluating the performance functions. Besides, the quantities that dominate the communication cost are $\sum_{k,l}$.

Please amend paragraph 96 to correct the symbols and the equation. In the equation, please note the distinction between the summation symbol and the subscripted variable Σ .

[0096] The l^{th} computing unit collects the SS, $\{n_{k,l}, \sum_{k,l}, s_{k,l} \mid k=1, \dots, K\}$, on the data in its own memory, and then sends them to the Integrator. The Integrator simply adds up the SS from each unit to get the global SS,

Best Available Copy

Appl. No. 09/844,730

Amdt. dated July 19, 2005

Reply to Notice of Allowance of July 29, 2005

$$n_k = \sum_{l=1}^L n_{k,l}, \quad \Sigma_k = \sum_{l=1}^L \Sigma_{k,l}, \quad s_k = \sum_{l=1}^L s_{k,l}$$

$$n_k = \sum_{l=1}^L n_{k,l}, \quad \Sigma_k = \sum_{l=1}^L \Sigma_{k,l}, \quad s_k = \sum_{l=1}^L s_{k,l}.$$

Please amend paragraph 97 to correct the symbols as follows:

[0097] The leading cost of integration is $O(K \cdot \dim \cdot L)$, where L is the number of computing units. The new location of the k^{th} center is given by $m_k = [\equiv] \Sigma_k / n_k$ from the global SS (this is the $I(\cdot)$ function in (2)), which is the only information all the computing units need to start the next iteration. The performance function is calculated by (proof by direct verification),

$$Perf_{KM} = \sum_{l=1}^L s_k.$$

Please amend paragraph 102 to correct the equation. Note that the numerator in the first summation has changed from "1" to "x".

[0102] (K-Means is similar, except its weights are the nearest-center membership functions, making its centers centroids of the cluster.) Overall then, the recursion equation is given by

$$m_k = \frac{\sum_{x \in S} \frac{1}{d_{x,k}^3 \left(\sum_{l=1}^K \frac{1}{d_{x,l}^2} \right)^2}}{\sum_{x \in S} \frac{1}{d_{x,k}^3 \left(\sum_{l=1}^K \frac{1}{d_{x,l}^2} \right)^2}}$$

$$m_k = \frac{\sum_{x \in S} \frac{x}{d_{x,k}^3 \left(\sum_{l=1}^K \frac{1}{d_{x,l}^2} \right)^2}}{\sum_{x \in S} \frac{1}{d_{x,k}^3 \left(\sum_{l=1}^K \frac{1}{d_{x,l}^2} \right)^2}} \quad (9)$$

Appl. No. 09/844,730
 Amdt. dated July 19, 2005
 Reply to Notice of Allowance of July 29, 2005

Please amend paragraph 103 to correct the equation. Note that the index for the summation is "k" rather than "λ", and the denominators in the expression for g₂ and g₃ are cubed "3" rather than taken to the S power "S".

[0103] where $d_{x,k} = \|x - m_k\|$ and s is a constant $\cong 4$. The decomposed functions for calculating SS (see (3) and (4)) are then

$$\left[\begin{array}{l} g_1(x, M) = 1 / \sum_{\lambda=1}^K \frac{1}{d_{x,k}^2} \\ g_2(x, M) = g_1^2(x, M) \cdot \left(\frac{1}{d_{x,1}^S}, \frac{1}{d_{x,2}^S}, \dots, \frac{1}{d_{x,K}^S} \right) \\ g_3(x, M) = g_1^2(x, M) \cdot \left(\frac{1}{d_{x,1}^S}, \frac{1}{d_{x,2}^S}, \dots, \frac{1}{d_{x,K}^S} \right) x \end{array} \right] \left[\begin{array}{l} g_1(x, M) = 1 / \sum_{k=1}^K \frac{1}{d_{x,k}^2} \\ g_2(x, M) = g_1^2(x, M) \cdot \left(\frac{1}{d_{x,1}^3}, \frac{1}{d_{x,2}^3}, \dots, \frac{1}{d_{x,K}^3} \right) \\ g_3(x, M) = g_1^2(x, M) \cdot \left(\frac{x}{d_{x,1}^3}, \frac{x}{d_{x,2}^3}, \dots, \frac{x}{d_{x,K}^3} \right) \end{array} \right]$$

Please amend paragraph 108 to correct the equation. Note that the subscript for p is "k" rather than "λ", and the symbol for summation must be carefully distinguished from the subscripted variable Σ.

[0108] In this example, the EM algorithm with linear mixing of K bell-shape (Gaussian) functions is described. Unlike K-Means and K-Harmonic Means in which only the centers are to be estimated, the EM algorithm estimates the centers, the co-variance matrices, Σ_k , and the mixing probabilities, $p(m_k)$. The performance function of the EM algorithm is

$$\text{Perf}_{EM}(X, M, \Sigma, p) = -\log \left\{ \prod_{x \in S} \left[\sum_{k=1}^K p_k \cdot \frac{1}{\sqrt{(2\pi)^D \det(\Sigma_k)}} \cdot \text{EXP} \left(- (x - m_k) \Sigma_k^{-1} (x - m_k)^T \right) \right] \right\}$$

$$\text{Perf}_{EM}(X, M, \Sigma, p) = -\log \left\{ \prod_{x \in S} \left[\sum_{k=1}^K p_k \cdot \frac{1}{\sqrt{(2\pi)^D \det(\Sigma_k)}} \cdot \text{EXP} \left(- (x - m_k) \Sigma_k^{-1} (x - m_k)^T \right) \right] \right\}$$

(13)

Best Available Copy

Appl. No. 09/844,730
 Amdt. dated July 19, 2005
 Reply to Notice of Allowance of July 29, 2005

Please amend paragraph 111 to correct the equation. Note that the symbol for summation must be carefully distinguished from the subscripted variable Σ_k .

[0111] where $p(x|m)$ is the prior probability with Gaussian distribution, and $p(m_k)$ is the mixing probability.

$$p(x|m_k) = \frac{1}{\sqrt{(2\pi)^D \det(\Sigma_k)}} \cdot \text{EXP} \left(- (x - m_k) \sum_k^{-1} (x - m_k)^T \right)$$

$$p(x|m_k) = \frac{1}{\sqrt{(2\pi)^D \det(\Sigma_k)}} \cdot \text{EXP} \left(- (x - m_k) \Sigma_k^{-1} (x - m_k)^T \right) \quad (15)$$

Please amend paragraph 112 to correct the equation. Note that the subscript for m is "k" rather than " λ ", and the symbol for summation must be carefully distinguished from the subscripted variable Σ .

[0112] M-Step: With the fuzzy membership function from the E-Step, find the new center locations, new co-variance matrices, and new mixing probabilities that maximize the performance function.

$$m_k = \frac{\sum_{x \in S} p(m_k|x) \cdot x}{\sum_{x \in S} p(m_k|x)}, \Sigma_k = \frac{\sum_{x \in S} p(m_k|x) \cdot (x - m_k)^T (x - m_k)}{\sum_{x \in S} p(m_k|x)}, p(m_k) = \frac{1}{|S|} \sum_{x \in S} p(m_k|x)$$

$$m_k = \frac{\sum_{x \in S} p(m_k|x) \cdot x}{\sum_{x \in S} p(m_k|x)}, \Sigma_k = \frac{\sum_{x \in S} p(m_k|x) \cdot (x - m_k)^T (x - m_k)}{\sum_{x \in S} p(m_k|x)}, p(m_k) = \frac{1}{|S|} \sum_{x \in S} p(m_k|x) \quad (16-18)$$

Best Available Copy

Appl. No. 09/844,730
 Amdt. dated July 19, 2005
 Reply to Notice of Allowance of July 29, 2005

Please amend paragraph 113 to correct the equation. Note that in the expression for f_1 , the subscript for m is "k" rather than " λ ".

[0113] The functions for calculating the SS are:

$$\begin{aligned}
 \cancel{f_1(x, M, \Sigma, p)} &= \cancel{-\log \left[\sum_{l=1}^K p(x | m_l) p(m_l) \right]} \\
 f_1(x, M, \Sigma, p) &= -\log \left[\sum_{l=1}^K p(x | m_l) p(m_l) \right] \\
 g_1(x, M, \Sigma, p) &= (p(m_1 | x), p(m_2 | x), \dots, p(m_K | x)) \\
 g_2(x, M, \Sigma, p) &= (p(m_1 | x)x, p(m_2 | x)x, \dots, p(m_K | x)x) \\
 g_3(x, M, \Sigma, p) &= (p(m_1 | x)x^T x, p(m_2 | x)x^T x, \dots, p(m_K | x)x^T x)
 \end{aligned}$$